SPPPred: Sequence-Based Protein-Peptide Binding Residue Prediction Using Genetic Programming and Ensemble Learning

Shima Shafiee, Abdolhossein Fathi, and Ghazaleh Taherzadeh

Abstract—Peptide-binding proteins play significant roles in various applications such as gene expression, metabolism, signal transmission, DNA (Deoxyribose Nucleic Acid) repair, and replication. Investigating the binding residues in protein-peptide complexes, especially from their sequence only, is challenging experimentally and computationally. Although several computational approaches have been introduced to determine and predict these binding residues, there is still ample room to improve the prediction performance. In this work, we introduce a novel ensemble machine learning-based approach called SPPPred (Sequence-based Protein-Peptide binding residue Prediction) to predict protein-peptide binding residues. First, we extract relevant sequential information and employ genetic programming algorithm for feature construction to find more distinctive features. We then, in the next step, build an ensemble-based machine learning classifier to predict binding residues. The proposed method shows consistent and comparable performance on both ten-fold cross-validation and independent test set. Furthermore, SPPPred yields F-Measure (F-M), Accuracy(ACC), and Matthews' Correlation Coefficient (MCC) of 0.310, 0.949, and 0.230 on the independent test set, respectively, which outperforms other competing methods by approximately up to 9% on the independent test set. SPPPred is publicly available.https://github.com/GTaherzadeh/SPPPred.git.

Index Terms—Binding residue prediction, ensemble learning, genetic programming, protein-peptide interaction, sequence-based

1 Introduction

Nucleic acids, proteins, carbohydrates, lipids, and peptides are usually identified as vital molecules and have various functions upon interaction [1], [2]. The interactions between proteins and peptides play critical roles in several applications such as biological functions, cancer treatment, signal cascades, regulatory networks, immune responses, and enzyme inhibition [1], [2], [3]. These interactions are often identified using various experimental approaches. However, experimental approaches are limited due to the small peptide sizes, short separated binding motifs, weak binding affinity, internal protein labeling, and expensive, laborious, and time-consuming analysis of the obtained spectra [2], [4], [5], [8]. In this way, computational approaches, including template-, machine learning-, and deep learning-based, are employed on the sequence or structure of proteins to assist experimental approaches. Traditionally, structural templates were known to be effective because of limitations such as the transient nature of protein-peptide interactions and peptide flexibility. Several template-based methods, such as PDBinder [4], SPOT-peptide [5], and GalaxyPepDock [6], were built to predict protein-peptide binding sites. Inter Pep [7], employed sequence information, structural template matches, hierarchical clustering, and the Random Forest (RF) algorithm to identify peptide-binding sites. Taherzadeh et al. [8] proposed a novel structure-based approach to predict peptide-binding interactions by applying the RF classifier and a clustering algorithm, respectively. Moreover, Shafiee and Fathi [9], confirmed a Genetic Programming(GP) feature construction, Support Vector Machine (SVM) classifier, and Ordering Points To Identify the Clustering Structure (OPTICS)clustering by employing predicted structure and sequence-based information. Although the prediction performance of structure-based methods is more accurate, they are limited to predicting the binding sites of structurally known complexes only. In 2016, the first sequence-based method was introduced by Taherzadeh et al. [10] for the prediction of protein-peptide residue-level interactions. In other studies, [11], [12], various machine learning-based techniques (the SVM classifier as well as the Extra Tree (ET) classifier and Bagging Classifier (BC)) for predicting protein-peptide binding residues by employing predicted structure- and sequence-based information were adopted. For the prediction of peptide-binding residues, a consensus-based (combination of sequence and two RF template-based) method was proposed [13]. Besides, machine-learning methods can be categorized as individual- and ensemble-based. RF was adopted as the ensemble-based framework to predict binding residues in both bound and unbound structures [14]. The stacking-based method with two-tier learning [15] and the ensemble-based method consisting of distinct learning algorithms [16] to detect residue-based patterns by sequence information was proposed. In the last few years, the deep learning (DL) computing paradigm

- *Shima Shafiee and Abdolhossein Fathi are with the Department of Computer Engineering and Information Technology, Razi University, Kermanshah 67144-14971, Iran. E-mail: {shafiee.shima, a.fathi}@razi.ac.ir.*
- *Ghazaleh Taherzadeh is with the Department of Mathematics and Computer Science, Wilkes University, Wilkes-Barre, PA 18766 USA. E-mail: ghazaleh.taherzadeh@wilkes.edu.*

Manuscript received 23 June 2022; revised 12 November 2022; accepted 10 December 2022. Date of publication 19 December 2022; date of current version 5 June 2023.
(Corresponding author: Abdolhossein Fathi.)
Digital Object Identifier no. 10.1109/TCBB.2022.3230540

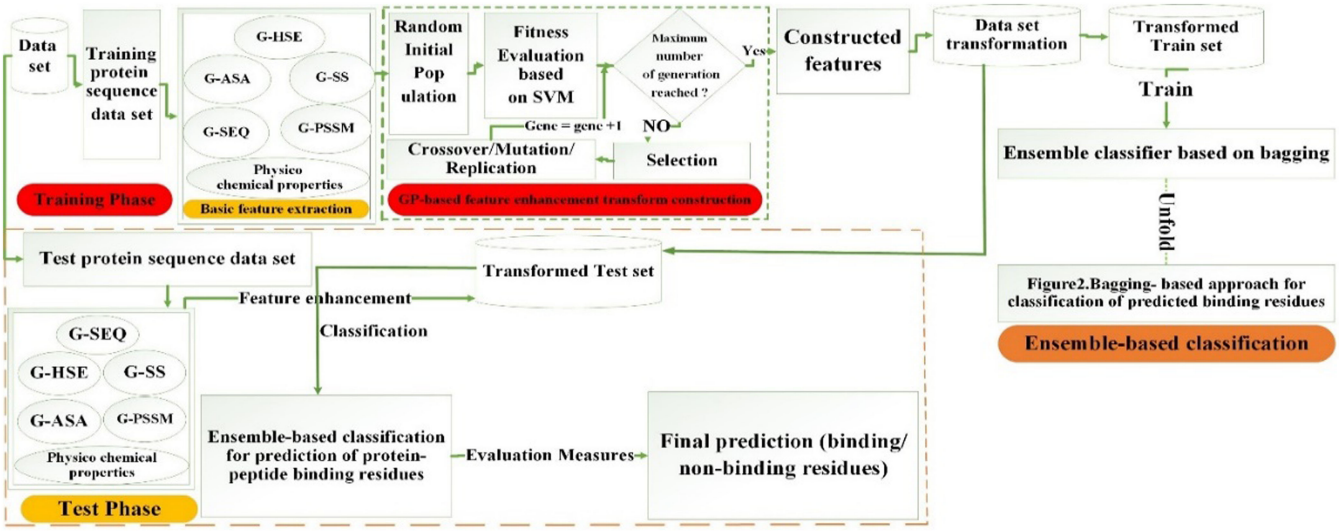

Fig. 1. The proposed architecture of SPPPred.

has been adopted as an effective technique in the machine learning (ML) community. For instance, Convolutional Neural Networks (CNNs) [17], [18] are widely applied in Sharma's s innovations. His team proposed CNN based on the transformation of non-image data to image ones. Their proposed frameworks were adopted by employing RNA-Seq, vowels, text, and artificial [17] as well as high-dimensional omics data and effective sets of genes [18]. In this way, a multiple-task deep learning approach [19] as well as, the deep-based predictors [20], [21], [22] were established to predict the protein-peptide interactions and protein-peptide binding regions in terms of sites and residues. Other deep learning-based models [23], [24] were developed for the prediction of peptide-binding regions using both the sequence and structure of proteins. However, the prediction performance of these deep learning algorithms highly depends on the number of very limited complexes.

As a result, our explicit objectives based on motivating questions are the following.

1. Are the constructed features more meaningful and distinctive high-level? How?
2. How is the effect of the GP-based feature construction techniques on the performance of the types of machine-learning-based classifiers including individual and ensemble-based?
3. Is the robustness of the proposed predictor justified in protein-peptide binding residue detection?
4. Is an interdisciplinary evaluation of the given bioinformatics case study?

Eventually, this paper is organized into the following sections. Section 2 provided the Materials and methods. In Section 3, experimental results and discussion are elaborated. Finally, Section 4 presents conclusions and future direction.

2 Materials and Methods

2.1 The Proposed Method

The focus of our Sequence-based Protein-Peptide binding residue Prediction (or SPPPred) method relies on a novel two-level framework including feature construction and ensemble-based classification. In other words, feature construction generates discriminative and new high-level features as well as discovering the hidden relations of constructing new high-level and low-level features [25], [26] which improves the prediction performance. Although the training phase of GP is time-consuming, it increases the accuracy of the classifier while significantly reducing the feature space dimension. Moreover, we are dealing with imbalanced data for the prediction of peptide-binding residues in proteins. In the protein sequence, the number of non-peptide-binding residues is about 17 times more than the number of peptide-binding residues which leads to a biased prediction. To address this problem, an ensemble learning approach is adopted. Therefore, the proposed method is the fusion of a bagging-based classifier along with GP-based feature construction for predicting the class of residues in a given protein sequence that binds with peptides. Fig. 1 illustrates the overall framework of SPPPred. According to it, at first, we extract the basic feature groups namely Half Sphere Exposure Feature Group (G-HSE), Secondary Structure Group (G-SS), Accessible Surface Area Group (G-ASA), Position-Specific Scoring Matrix Group (G-PSSM), Sequence Group (G-SEQ), and physicochemical properties from input protein sequences. Next, the GP-based feature enhancement technique is applied for feature construction based on all feasible combinations of basic (unenhanced) features. Furthermore, dataset transformation with constructed (or enhanced) features is converted into new datasets (or transformed datasets) including transformed independent tests and transformed training sets. Finally, the ensemble-based classification is trained by using the transformed features of the training set without balancing binding and non-binding residues.

In the test phase, basic (or primary) features such as G-HSE, G-SS, G-ASA, G-PSSM, G-SEQ, and physicochemical properties are extracted from the input protein sequence. Next, enhanced features are obtained through the proposed GP-based feature enhancement and then the ensemble-based binary prediction (binding or non-binding residues) is employed.

2.2 Basic Features Extraction

Since the structures of several proteins for our dataset are not yet solved. According to our investigation, sequence evolution, as well as predicted structural information was extracted. Subsequently, various discriminative features are categorized into six groups based on their characteristics from the protein sequence.

2.2.1 G-HSE

Solvent exposure of amino acid residues presented significant information for investigating and predicting protein function, interactions, and structure [27]. G-HSE is a measure of protein solvent exposure that describes how buried amino acid residues are in the protein. Half-sphere exposure [27] is calculated using the Contact Numbers (CN) in upward and downward hemispheres along with a pseudo-Cβ-Cα vector. Cβ-Cα vector is calculated by dividing the contact number sphere into two halves using the perpendicular plane to the Cβ-Cα vector. This division of the contact number sphere produces HSE-up and HSE-down [27]. In this study, SPIDER3 [28] is used to obtain this feature group.

2.2.2 G-SS

The Secondary Structure (SS) represents the three-dimensional conformation of local segments of the protein [10]. Predicted SS values are SS probabilities for the type of classes such as α-helix, β-sheet, and coil. These predicted values are obtained using SPIDER 2.0 [29].

2.2.3 G-ASA

The Accessible Surface Area (ASA) refers to the surface area of a residue within a protein that is accessible to a solvent. The ASA values are obtained by SPIDER 2.0 [29] and transformed into the relative ASA (or rASA) values [10]. Then, the average rASA of adjacent amino acid residues is calculated within window sizes ranging from one to the optimal window size value (or rASA-avg) [10].

2.2.4 G-PSSM

The Position-Specific Scoring Matrix (PSSM) is the type of evolutionary-based information scheme that is commonly employed for the pattern representation of proteins. This representation is captured based on a category of sequences formerly aligned by using sequence or structural similarity [30]. PSSM is a matrix of L $\times$ 20 dimensions, where L is the protein length.

$$Protein\ pssm = \begin{bmatrix} E1 \rightarrow 1 & \dots E1 \rightarrow k \dots & E1 \rightarrow 20 \\ \vdots & \vdots & \vdots \\ Ei \rightarrow 1 & \cdots Ei \rightarrow k \dots & Ei \rightarrow 20 \\ \vdots & \vdots & \vdots \\ El \rightarrow 1 & \dots El \rightarrow k \dots & El \rightarrow 20 \end{bmatrix}$$

Columns indicate the 20 standard amino acid types ($k = 1\ldots20$) and each row in PSSM represents an amino acid in the given protein sequence. The sequence profiles are obtained using PSI-BLAST [31] by setting the cut-off E-value to 0.001 with three iterations [32].

2.2.5 G-SEQ

A 20-dimensional binary vector represents the position of each amino acid in the protein sequence by encoding the amino acid to 1 or 0. The type of amino acid residue at the sequence position is encoded by 1 otherwise 0 [33].

2.2.6 Physicochemical Properties

Each amino acid type has different physicochemical characteristics. In this study, steric parameters, hydrophobicity, volume, polarizability, isoelectric point, helix, and sheet probability physicochemical properties are used [10] and extracted by using the DisPredict2 [34]. We would like to note that, all feature groups are normalized using

$$x' = {(x - \min(x))}/{(\max(x) - \min(x))} \quad (1)$$

Where, x and x' are defined as the original and normalized values of the basic feature, respectively.

2.3 GP-Based Feature Construction

Recently, Evolutionary Computation (EC), like GP, has been extensively employed in feature construction [35] to construct new high-level features [36] as well as generate complex pattern representations such as trees, mathematical models with different operators, and functions automatically [37]. Likewise, the GP's procedure is depicted in Algorithm 1. Subsequently, this procedure contains generating an initial population of programs, evaluating computer programs, selecting, and evolving [38]. The proposed feature construction system (Fig. 2) is illustrated by constructing more meaning according to the hidden relations of low-level features as well as performing a more accurate classification task (Refer to Tables 5 and 6).

2.4 Ensemble Learning-Based Classification

Ensemble learning is one of the well-known approaches to controlling the imbalanced dataset [39], allowing a set of models to vote, achieving better performance and enhancing the predictive model, and reducing variance to improve generalization [40], [41].

In this study, bagging [42], is employed as an ensemble learning-based classification to predict peptide-binding residues (See Algorithm 2). Consequently, Fig. 3 also describes the bagging's steps that involve such as 1) bootstrapping feature randomize, which contains generating multiple datasets through random sampling with replacement. 2) prediction based on parallel training, which contains constructing multiple learning-based models (or classifiers) in parallel. Each Prediction Modeli (PMi), is trained by using a related Subset Data Setl (SDSl). 3) aggregation, which contains majority-voting outputs. The class with the majority votes is identified as the final predicted binding residue.

3 Experimental Results and Discussion

3.1 Dataset

To train and evaluate the proposed method, the initial protein-peptide dataset was extracted from BioLip [43] with the following conditions [20]: 1) Peptides are identified as chains with less than thirty amino acid residues. 2) Proteins

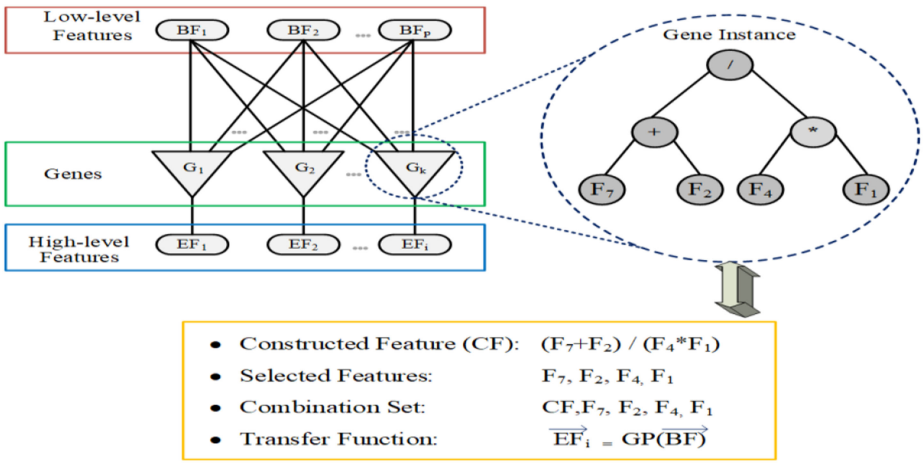

Fig. 2. The demonstration of GP-based feature construction.

with more than 30% similarity are excluded using Blastclust [44]. Subsequently, the final dataset [8], [20] consists of 1241 protein-peptide complexes with 16678 binding amino acid residues, and 280920 non-binding amino acid residues out of a total of 297598 amino acid residues. We randomly considered 10% of protein complexes as a test set and 90% as a training set for independent evaluation as performed in [8], [20]. The training set contains 1116 proteins including 14959 binding amino acid residues and 251769 non-binding amino acid residues out of a total of 266728 amino acid residues. The independent test set contains 125 proteins including 1719 binding amino acid residues and 29151 non-binding amino acid residues out of a total of 30870 amino acid residues. We also evaluated the proposed model using the ten-fold cross-validation to prove the robustness and consistency of SPPPred performance. To further understand, we analyzed the percentage of actual amino acids and the number of different actual residues types (actual binding or non-binding residues) for each amino acid on both the ten-fold cross-validation and independent test datasets. The experimental results are revealed in Fig. 4. Besides, Fig. 4a. shows the percentage of actual amino acids in both the independent test set and ten-fold cross-validation datasets which prove enrichment in amino acids including Alanine (A), Leucine (L), Valine(V), and Glutamic Acid(E). Figs. 4b and 4c depict non-binding residues enriched in A, L, V, and E for both mentioned datasets. As a result, all bar plots can confirm a class ratio of approximately 1:17 and subsequently the urgency of the ensemble learning-based technique for addressing the used imbalanced dataset.

3.2 Performance Evaluation Metrics

In Table 1, different metrics were adopted to evaluate the performance of SPPPred as well as to compare it with the competing methods. F-M is presented as a tradeoff between True Positive Rate (TPR) and precision. Therefore, Precision-Recall (PR) based criteria give a more informative view of a model's performance when there exist imbalanced datasets [45]. In highly imbalanced problems, PR-based metrics normalize the number of false positives concerning the number of true negatives whereas precision-based metrics normalize it according to the number of true positives. Specificity (SPE) and Sensitivity (SEN) explain the Accuracy (ACC) of a test set mathematically that describes the absence or presence of a condition compared to a definition. SEN is related to how well the actual binding residues are noticed as binding residues [46]. Mathew's Correlation Coefficient (MCC) is the type of statistical rate that

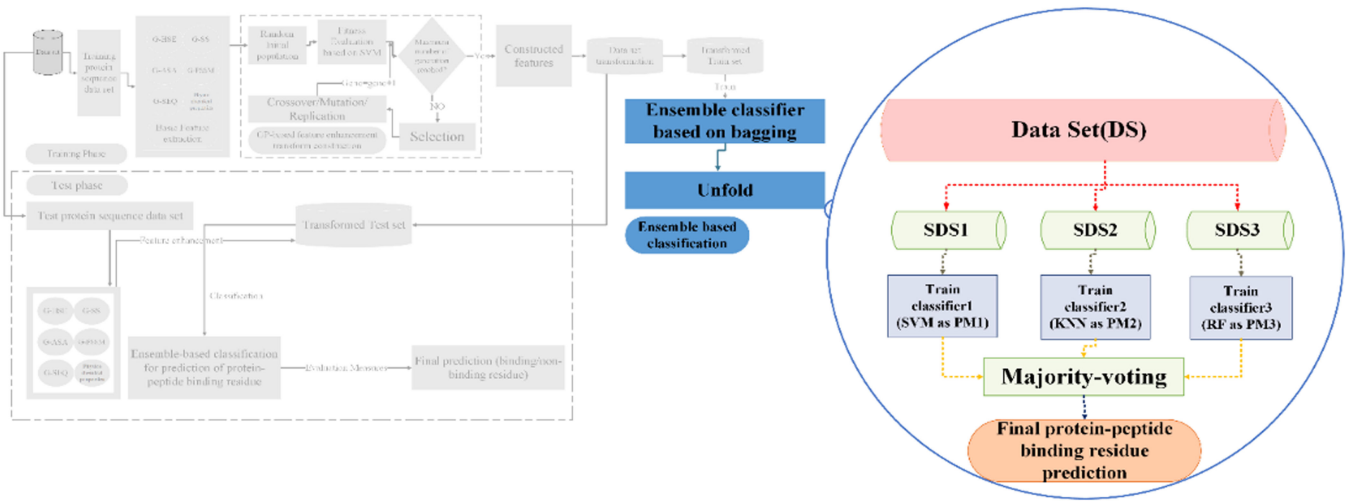

Fig. 3. The general block diagram of the bagging-based classifier.

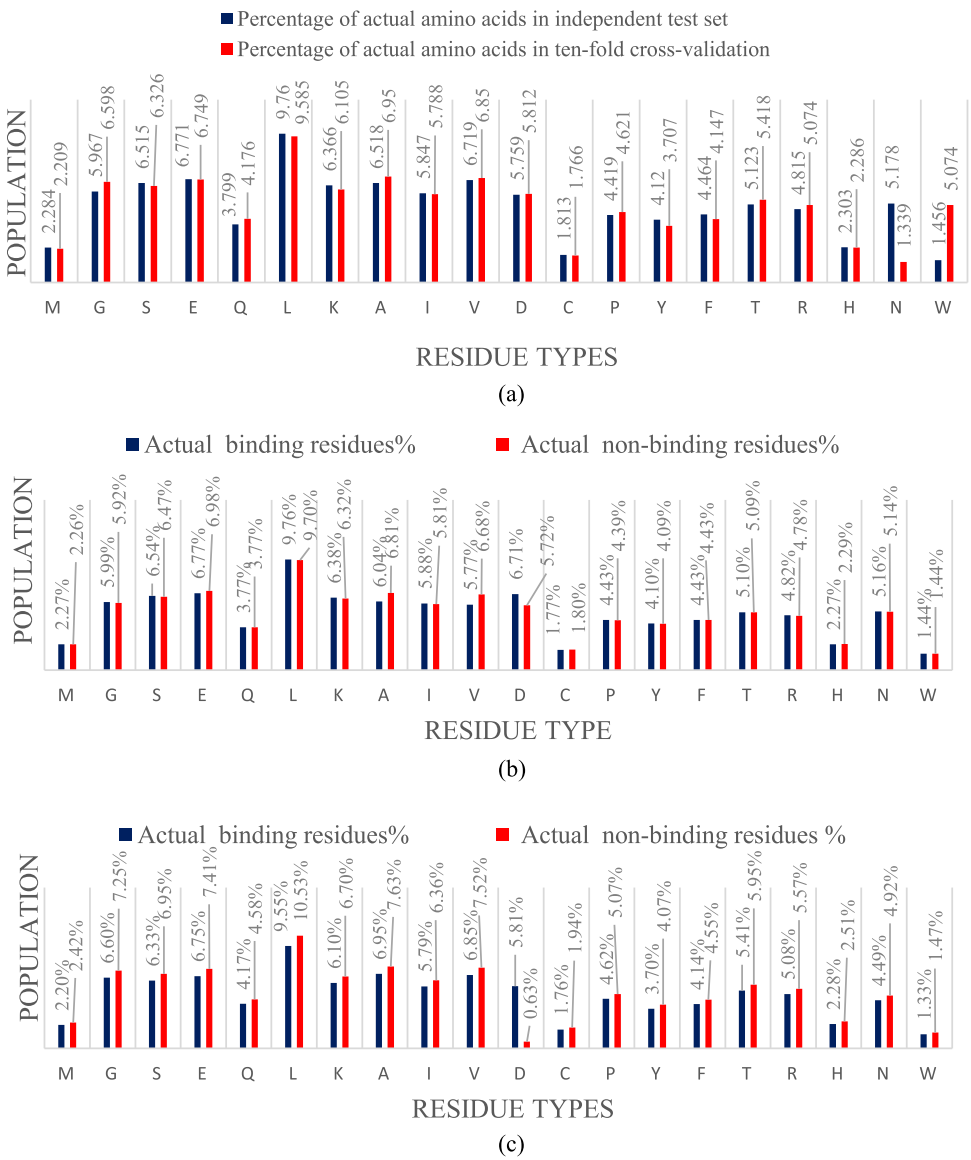

Fig. 4. By considering the defined class rate on, (a) the percentage of the population distribution of actual amino acid binding and non-binding residues according to residue types on (b) the independent test set, and (c) the ten-fold cross-validation.

TABLE 1
Name and Definition of the Evaluation Criteria

Metric	Description
False Negative(FN)	Incorrectly predicted non-peptide binding
False Positive (FP)	Incorrectly predicted peptide binding residues
True Negative (TN)	Correctly predicted non-peptide binding
True Positive (TP)	Correctly predicted peptide binding residues
Area Under Curve(AUC)	The area under the receiver operating characteristics (ROC) curve
Receiver operating characteristics (ROC)	A curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR)
Accuracy (ACC)	$(\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN})$
Sensitivity (SEN) or True Positive Rate (TPR)	$\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$
Specificity(SPE)) or True Negative Rate (TNR)	$\mathrm{TN}/(\mathrm{FP}+\mathrm{TN})$
False Positive Rate (FPR)	$\mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$
F − Measure(F-M) or (Harmonic mean of precision and recall)	$2*\mathrm{TP}/(2*\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$
Mathew's Correlation Coefficient(MCC)	$(\mathrm{TP}*\mathrm{TN})-(\mathrm{FP}*\mathrm{FN})/\sqrt{(\mathrm{TP}+\mathrm{FP})*(\mathrm{TP}+\mathrm{FN})*(\mathrm{TN}+\mathrm{FP})*(TN+FN)}$

TABLE 2
The Details of the Proposed Genetic Programming Algorithm According to Applied Mathematical Operations

Name of operator	Description	Name of operator	Description
SQRT	The square root of a value	EQA	Operator a
A tan	The arctangent of a value	ABS	Absolute operator
EXP	Exponential a value	HAT	Hat function of operator a
SIN	Sine operator in radians	Max	Maximum
COS	Cosine operator in radians	A cos	Arc cosine of a value
SOFT PLUS	$0.2 \log 1.0 + exp\ a$	A sin	Arcsine of a value
SQUARE	Power tow of a value	CONST	Constant value output
MUL	Multiplication operator	LOG	Logarithm
DIV	Division operator	NEG	Negative
CUBE	The third power of the operator	ADD	Addition operator
Quo	Quotient	SUB	Subtraction operator

generates a computed score that relies on prediction when the impact of non-binding residues is as significant as binding residues [47].

Algorithm 1. The Framework of the Proposed GP

Input: The terminal set, arithmetic operation, and parameter setting
Output: Return the best chromosome as solutions (feature construction functions)
Begin
1. **Generate** an initial random population P at iteration zero by using the Ramped half-and-half method;
2. P ← 0
3. g ← 0
4. **While** (P < the maximum number of generations and random chromosome C_i is not a member of the initial population P) **do**
5. **For** each GP tree:
6. **Get** the Subtrees of GP
7. **Convert** datasets
8. **Evaluate** the converted datasets based on SVM with RBF kernel
9. **Assign** the fitness Value
10. **End while**
11. **For** i:1 to N do
12. **If** (Ci is not a member of the initial population P) **then** P← the union of chromosome set P and random Chromosome Ci in the unique collection
13. **End for**
14. EP ← Compute fitness of P
15. G ← 1
16. **While** ((maximum fitness) and (final generation)) **do**
17. **Conduct** selection to choose chromosomes
18. **Conduct** crossover, mutation, and reproduction operations based on the selected chromosomes
19. **Update** population somehow to get C new genes
20. EP ← Compute fitness of P
21. G ← G + 1
22. **End while**
23. **Return** the best chromosome with the best-so-far fitness value

End

The Area Under the Curve (AUC) is the measure of the capability of the classifier to identify between the positive and negative classes [48], [49]. In general, AUC values are existence in the interval of [0, [1], when significant disparities are existing in the value of false positives vs. false negatives. So, minimizing one classification error such as a false negative is the critical item [49], [50]. ACC is a type of metric that shows the proportion of true outcomes amongst the whole number of instances tested. It is an effective classification metric for binary classification problems [47], [51].

3.3 Parameters Setting and Evaluation

The proposed GP-based feature construction employed mathematical operators and various executive parameters are described in Tables 2 and 3.

Note, to select the optimal kernel function, various kernel functions, including [49] Linear, Poly, and Radial Basis Function (RBF) were evaluated. The obtained results confirmed [9] the optimality of the RBF kernel that is determined in equation 2.

$$k(x, x') = \exp\left(\frac{-||x1 - x2||^2}{2\sigma^2}\right) \tag{2}$$

$||x1 - x2||^2$can be identified as the squared Euclidean distance between the two feature vectors in each dimensional space and $\boldsymbol{\sigma}$ is a free parameter.

TABLE 3
Parameter Settings for the Proposed Genetic Programming

Name of parameters	Description
Number of generations	500
Population Size	500
Cross Over Strategy	Uniform
Cross Over Rate	0. 800
Mutation Strategy	Mutation
Mutation Rate	0.300
Selection Strategy	Tournament
Size of Tournament	9
Max Tree Depth	12
Competition size	8
Reproduction Rate	0.109
Minimum initial tree size	5
Maximum initial tree size	10
Terminal condition	Max generation
Constants Leave Probability	0.100

TABLE 4
Performance of Individual-Based Classifiers on the Ten-Fold Cross-Validation

Machine learning algorithm/ Metric	MCC	AUC	F-M	ACC	SEN	SPE
SVM	**0.130±0.019**	**0.549±0.049**	**0.177±0.025**	**0.738±0.074**	**0.208±0.037**	**0.789±0.089**
KNN	0.128±0.020	0.542±0.050	0.170±0.028	0.732±0.077	0.204±0.039	0.787±0.092
RF	0.124±0.023	0.540±0.055	0.168±0.030	0.731±0.080	0.201±0.042	0.786±0.093
NB	0.109±0.030	0.509±0.060	0.148±0.035	0.719±0.095	0.189±0.059	0.749±0.110
LR	0.100±0.033	0.504±0.060	0.146±0.037	0.718±0.097	0.188±0.060	0.743±0.112

Algorithm 2. The Framework of Bagging

Input:
1. Using correct labels $\alpha_i \in \theta = \{\alpha_1, \ldots, \alpha_C\}$ representing K classes of the problem for training data SDSn
2. Using weak machine learning-based algorithms as weak classifiers
3. Considering L as the number of iterations
4. Generating bootstrapped training data based on percent P

Output:
1. Composite model

Method:

Train phase:

For t = 1, . . ., L, **do**
1. Take a bootstrapped sample SDS from the dataset, by using randomly drawing the P percent of SDS
2. Call weak learn using SDS_l and capture the hypothesis (classifier) b_l.
3. Considering E as the ensemble via appending b_l to it

End For

End Train phase

Evaluation phase:

Begin
1. Evaluate the ensemble $E = \{b_1, \ldots, b_L\}$ on Y | | Y as given unlabeled instances Y
2. Let $MV_{t,g} = \begin{cases} 0, & otherwise \\ 1, & if\ b_l\ picks\ class\ \alpha_i \end{cases}$

Assign the vote given to the class α_i by using classifier b_l.
3. Receive the whole vote received by using every class $MV_g = \sum_l^L MV_{l.g}, g = 1, \ldots, K$
4. Terminal classification is obtained via the class with the highest whole vote

End

End Evaluation phase

End Method

According to the number of peptide-binding residues is significantly lower than the number of non-peptide-binding residues, the standard deviation (or StdDev) of ten trials was also calculated (See Table 4). In this way, to evaluate the proposed predictor called the bagging framework, five different machine learning algorithms (SVM [52], RF [53], K-Nearest Neighbor (KNN [54]), Naive Bayes (NB) [55], and Linear Regression (LR) [56]) were evaluated by using the ten-cross-validation and the independent test set.

The obtained results in Tables 4 and 5 show that SVM performed better than the other classifiers (KNN, LR, NB, and RF) on the independent test set and obtained the best performance with optimal StdDev by the ten-fold cross-validation. Next, we applied a grid search to determine the optimized number of residues around a target residue (window size), which can moderate the interaction between protein and peptide and establish an effective predictor. Subsequently, five machine learning algorithms (SVM, KNN, RF, NB, and LR classifiers) with five different window sizes (1, 3, 5, 7, and 9) were adopted.

The window size for which the classifier yields the highest performance on the used dataset was selected as the best window size for that classifier. In addition, Fig. 5 illustrates the optimal window sizes 31,97, and 5 for SVM, KNN, RF, NB, and LR on the independent test set, respectively. It means that the optimal window size for different classifiers is various which indicates the nature of employed classifiers differs from one another.

To select the best combination, we examined the performance of different combinations of single, triple, and quintuplets of the base classifiers with their optimal window sizes. Based on the obtained results in Table 6, for the single classifier, the SVM has higher performance on all metrics compared to others. In addition, employing the ensemble technique of SVM, KNN and RF (GP+ (SVM, KNN, RF)) has the highest performance on all metrics on the ten-fold cross-validation and independent test set. Therefore, the combination of GP+ (SVM, KNN, RF) is adopted as the final combination of the proposed method. Furthermore, by comparing the performance of single classifiers along with the GP approach and the performance of them without GP, shown in Tables 5 and 6, it is clear that employing GP-based feature enhancement increase the performance of all classifiers.

TABLE 5
Performance of Individual-Based Classifiers on the Independent Test Set

Machine learning algorithm/Metric	MCC	AUC	F-M	ACC	SEN	SPE
SVM [49]	**0.149**	**0.560**	**0.199**	**0.756**	**0.220**	**0.800**
KNN[51]	0.140	0.551	0.190	0.744	0.219	0.790
RF[50]	0.139	0.550	0.189	0.739	0.215	0.788
NB [52]	0.111	0.522	0.169	0.700	0.180	0.766
LR[53]	0.110	0.520	0.165	0.699	0.179	0.760

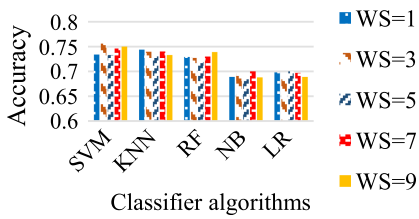

Fig. 5. Performance of optimal window sizes used in different classifiers namely, SVM, KNN, RF, NB, and LR on the independent test.

TABLE 6
The Comparison of SPPPred Performance on the Independent Test Set Using Different Classifiers

Combination	MCC	AUC	F-M	ACC	SEN	SPE
GP+SVM	0.202	0.699	0.240	0.930	0.279	0.940
GP+KNN	0.191	0.680	0.231	0.890	0.260	0.928
GP+RF	0.187	0.669	0.229	0.879	0.256	0.909
GP+NB	0.146	0.632	0.210	0.840	0.209	0.890
GP+LR	0.141	0.630	0.208	0.835	0.200	0.888
GP+(SVM,KNN,RF)	**0.230**	**0.710**	**0.310**	**0.949**	**0.315**	**0.959**
GP+(SVM,KNN,NB)	0.220	0.708	0.305	0.936	0.297	0.948
GP+(SVM,KNN,LR)	0.218	0.707	0.301	0.931	0.291	0.946
GP+(SVM,RF,NB)	0.215	0.705	0.300	0.929	0.290	0.944
GP+(SVM,RF,LR)	0.214	0.704	0.299	0.928	0.289	0.942
GP+(KNN,RF,NB)	0.190	0.680	0.280	0.909	0.279	0.929
GP+(KNN,RF,LR)	0.189	0.679	0.279	0.907	0.278	0.918
GP+(KNN,NB,LR)	0.188	0.678	0.277	0.905	0.275	0.916
GP+(RF,NB,LR)	0.169	0.669	0.259	0.889	0.269	0.899
GP+(SVM,KNN,RF,NB,LR)	0.218	0.707	0.299	0.928	0.287	0.941

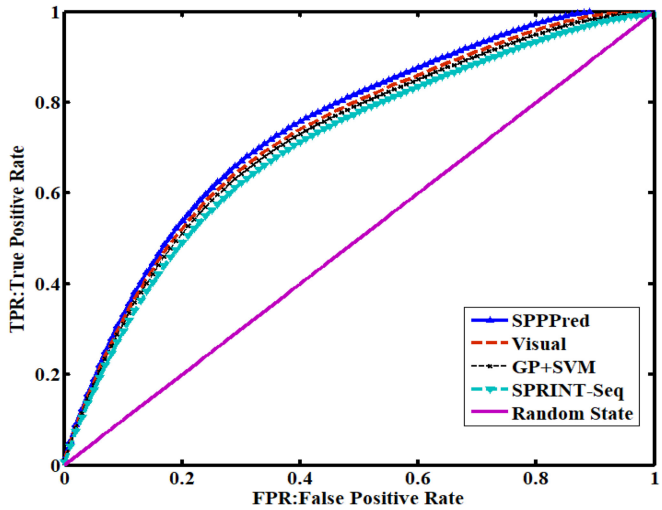

Fig. 6. ROC curves to show and compare the performance of SPPPred and other existing methods on the independent test set.

Furthermore, we compared the performance of the proposed machine learning-based predictor, called SPPPred, on the ten-fold cross-validation and independent test in Table 7. As a result, the comparative analysis in Table 7 indicates that the prediction performance on the independent test set is approximately 2% to 4.5% higher than the prediction performance on the ten-fold cross-validation. The small differences between the prediction performance of SPPPred on the ten-fold cross-validation and independent test set prove the robustness and generality of the proposed method.

Moreover, the achieved AUC values of the SPPPred method are 0.710 and 0.669 on the independent test set and ten-fold cross-validation, respectively. Moreover, Fig. 6 shows the comparison of four ROC curves indicating the performance of SPPPred, Visual [20], SVM+GP [9], and SPRINT-Seq [10] on the independent test set which proves the optimal consistency of SPPPred performance.

3.4 Comparison With Other Methods

Table 8 compares the performance of SPPPred and other existing works including SPRINT-Seq [10], GP+SVM [9], and Visual [20] on the independent test set of 125 proteins.

According to Table 8, the performance evaluations of the SPPPred for MCC, AUC, F-M, ACC, SEN, and SPE metrics are 0.230, 0.710, 0.310, 0.949, 0.315, and 0.959, respectively on the independent test set. It can be concluded that the proposed method can predict the type of class (binding or non-binding residues) for every residue with the best F-M, MCC, and ACC compared to competing approaches. Additionally, all performance evaluations are higher than the previous study by Shafiee and Fathi [9], called GP+SVM, with MCC = 0.202, AUC = 0.699, F-M = 0.240, ACC = 0.930, SEN = 0.279, and SPE = 0.940. According to the achieved SPE, 0.959 of actual non-binding residues are correctly classified as non-binding residues. The remaining percentage ($\sim$ 0.041) of actual non-binding residues is incorrectly classified as binding residues. Our method achieves the SPE value of 0.959, which is higher than all other existing methods except SPRINT-Seq [10]. Moreover, the proposed method achieves the SEN of 0.315 and the remaining percentage ($\sim$ 0.685) of actual binding residues is incorrectly classified as non-binding residues. In Table 8 the Visual method [20] achieves the highest SEN (0.670) on the independent test set. However, non-binding residue prediction is as important as binding residue prediction in our method. Thus, we focus on the F-M, which is not reported by Visual [20]. Table 8 also indicates high ACC for all mentioned methods due to an imbalanced dataset. Due to the nature of an imbalanced dataset, it is not surprising to achieve high ACC by predicting each test observation as the majority class. As explained, ACC cannot be considered a fair performance evaluation metric for such a binary classification problem.

However, SPPPred still achieves the best ACC of 0.949 on the independent test set. Although SPPPred achieves a slightly lower ($\sim$0.02) AUC value (0.710) compare to Visual = 0.730, but still close enough to be comparable. In this way, to visual-based represent the output of SPPPred, we employed 1dpu, chain A protein containing 69 residues with 11 actual binding residues. As illustrated, the predicted protein-peptide binding

TABLE 7
The Comparison of SPPPred Performance on the Ten-Fold Cross-Validation and the Independent Test Set

Dataset	MCC	AUC	F-M	ACC	SEN	SPE
Ten-fold cross-validation	0.190	0.669	0.289	0.929	0.279	0.925
Independent test set	**0.230**	**0.710**	**0.310**	**0.949**	**0.315**	**0.959**

TABLE 8
Comparison of the Performance of SPPPred and Existing Methods on the Independent Test Set

Methods/Metrics	MCC	AUC	F-M	ACC	SEN	SPE
SPRINT-Seq[10]	0.200	0.680	0.221	0.920	0.210	**0.960**
Visual[20]	0.170	**0.730**	-	-	**0.670**	0.680
GP+SVM[9]	0.202	0.699	0.240	0.930	0.279	0.940
SPPPred (our method)	**0.230**	0.710	**0.310**	**0.949**	0.315	0.959

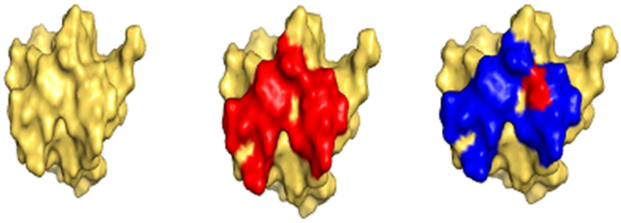

Fig. 7. (a) Original protein structure, PDB id: 1dpu, chain A. (b) Actual binding residues (red). (c) Predicted binding residues by SPPPred (blue).

residues by SPPPred (7. c) are very close to the actual ones (7. b). In Fig. 7, we should emphasize that SPPPred can predict 10 residues (blue) correctly of 11 actual binding residues (red) as well as it can be compared with the actual state.

In addition, the protein-peptide binding residues predicted by SPPPred are comparable to state-of-art existing methods (see Fig. 8).

To visual-based comparison of one-dimensional (sequence)-based representation was employed. Likewise, the comparison results of SPPPred and other competing methods, including experimental(actual) and computational states (GP+SVM, and Visual), were compared. It means that the proposed method can predict more true positive and fewer false positive rates. In Fig. 8, actual and predicted binding residues are indicated in red and blue colors, respectively. The proposed method can predict 21 residues correctly out of 24 actual binding residues out of 361 total residues and only has 5 false predicted binding residues. The Visual method [20] can predict 17 residues correctly out of 24 actual binding residues and also predicted 17 false binding residues. The GP+SVM [9] method can predict 19 residues correctly out of 24 actual binding residues and predict 6 false binding residues. Therefore, according to the comparison, SPPPred can predict more actual binding residues and has lower false binding residue prediction.

3.5 Computational Complexity Evaluation

The computational complexity is investigated to achieve a deep understanding of the GP's effect in reducing the dimension of feature space, constructing features automatically, as well as improving the F-M of the classifier. By comparison, SPPPred outperforms other competing approaches (SVM [9], GP+SVM [9]), with the highest F-M value of 0.310 and 0.29 on the independent test set as well as the ten-fold cross-validation, respectively. In the ten-fold cross-validation and the independent test set, SPPPred obtains a feature space dimension of 10 which is better than the 12 and 33 achieved by GP+SVM [9], and SVM [9]. Likewise, the time complexity of 799.872 ms, 38866.207 ms, and 91995.000 ms on the ten-fold cross-validation, as well as the time complexity of 777.871 ms, 34566.201 ms, and 89001.300 ms on the independent test set are obtained respectively by competition methods including SVM [9], GP+SVM [9], and SPPPerd. According to Table 9, we further observe that when F-M is increased then time complexity is decreased by using the independent test set. Finally, it can be concluded that the high time complexity of GP is paid off by reducing and constructing powerful and effective features. This high-time complexity occurs only once in the process of building the feature space but results in a consistent and higher performance of the final method.

3.6 Statistic-Based Evaluation

This paper addresses the application of two types of statistical tests through an evaluation question. Which can the method be more significant when GP-based feature construction was adopted? The statistic-based tests were employed to confirm statistical significance. We calculated

Actual (a)	KERLLDELTLEGVARYMQSERCRRVICLVGAGISTSAGIPDFRSYDNLEKYHLPYPEAIFEISYF	55-125
GP+SVM [9] (b)	KERLLDELTLEGVARYMQSERCRRVICLVGAGISTSAGIPDFRSYDNLEKYHLPYPEAIFEISYF	
Visual [20] (c)	KERLLDELTLEGVARYMQSERCRRVICLVGAGISTSAGIPDFRSYDNLEKYHLPYPEAIFEISYF	
SPPPred (d)	KERLLDELTLEGVARYMQSERCRRVICLVGAGISTSAGIPDFRSYDNLEKYHLPYPEAIFEISYF	
Actual (a)	KKHPEPFFALAKELYPGQFKPTICHYFMRLLKDKGLLLRCYTQNIDTLERIAGLEQEDLVEAHGTFYTSH	126-196
GP+SVM [9] (b)	KKHPEPFFALAKELYPGQFKPTICHYFMRLLKDKGLLLRCYTQNIDTLERIAGLEQEDLVEAHGTFYTSH	
Visual [20] (c)	KKHPEPFFALAKELYPGQFKPTICHYFMRLLKDKGLLLRCYTQNIDTLERIAGLEQEDLVEAHGTFYTSH	
SPPPred (d)	KKHPEPFFALAKELYPGQFKPTICHYFMRLLKDKGLLLRCYTQNIDTLERIAGLEQEDLVEAHGTFYTSH	
Actual (a)	CVSASCRHEYPLSWMKEKIFSEVTPKCEDCQSLVKPDIVFFGESLPARFFSCMQSDFLKVDLLLVMGTSL	197-267
GP+SVM [9] (b)	CVSASCRHEYPLSWMKEKIFSEVTPKCEDCQSLVKPDIVFFGESLPARFFSCMQSDFLKVDLLLVMGTSL	
Visual [20] (c)	CVSASCRHEYPLSWMKEKIFSEVTPKCEDCQSLVKPDIVFFGESLPARFFSCMQSDFLKVDLLLVMGTSL	
SPPPred (d)	CVSASCRHEYPLSWMKEKIFSEVTPKCEDCQSLVKPDIVFFGESLPARFFSCMQSDFLKVDLLLVMGTSL	
Actual (a)	QVQPFASLISKAPLSTPRLLINKEKAGQSDPFLGMIMGLGGGMDFDSKKAYRDVAWLGECDQGCLALAEL	268-338
GP+SVM [9] (b)	QVQPFASLISKAPLSTPRLLINKEKAGQSDPFLGMIMGLGGGMDFDSKKAYRDVAWLGECDQGCLALAEL	
Visual [20] (c)	QVQPFASLISKAPLSTPRLLINKEKAGQSDPFLGMIMGLGGGMDFDSKKAYRDVAWLGECDQGCLALAEL	
SPPPred (d)	QVQPFASLISKAPLSTPRLLINKEKAGQSDPFLGMIMGLGGGMDFDSKKAYRDVAWLGECDQGCLALAEL	
Actual (a)	LGWKKELEDLVRREHASIDAQS	339-361
GP+SVM [9] (b)	LGWKKELEDLVRREHASIDAQS	
Visual [20] (c)	LGWKKELEDLVRREHASIDAQS	
SPPPred (d)	LGWKKELEDLVRREHASIDAQS	

Fig. 8. The sequence-based schematic to the comparison of actual binding residues and predicted residues obtained by the proposed SPPPred method along with GP+SVM [9] and Visual method [20] on 4l3o protein sequence chain A (PDB id: 4l3oA). These sequences have been divided into sections (55-125, 126-196, 197-267268-338339-361) to fit the page. The upper row illustrates (a) the actual binding amino acid residues of the protein sequence highlighted with red color, while the lower rows illustrate the peptide-binding residues (Blue) as predicted by using competition approaches, (b) the GP+SVM approach [9], (c) Visual approach [20], (d) SPPPred (our proposed approach).

TABLE 9
Evaluation of Computational Complexity Using the Different Methods on the Ten-Fold Cross-Validation and Independent Test Set

Algorithms	Feature space dimension	Ten-fold cross-validation		Independent test set	
		F-M	Time (ms)	F-M	Time (ms)
SVM [9]	33	0.177	799.872	0.199	777.871
GP+SVM [9]	12	0.227	38866.207	0.240	34566.201
SPPPred	10	**0.289**	91995.000	**0.310**	89001.300

TABLE 10
Calculating a Two-Tailed Statistical T-Test of SPPPred and the Competing Methods

Method/Statistical-based evaluation	T-test	p-value
SPRINT-Seq [10]	30.999	0.038
Visual [20]	32.000	0.035
GP+ SVM [9]	33.111	0.032
SPPPred(our method)	**38.088**	**0.025**

statistic-based evaluation and the occurred results are presented in Table 10.

In Table 10, statistical analyses were performed by applying a significance cutoff of 0.05. Accordingly, the p-value is less than the defined threshold of 0.05 which can lead to a conclusion with significant differences. These differences exist between SPPPred and other competing methods (SPRINT-Seq [10], Visual [20], and GP+ SVM [9]). The above results show that SPPPred's results are the most significant over other mentioned methods.

4 Conclusion and Future Direction

In a brief, the most significant highlights of this study are as follows. (i)This work is the first application of GP-based feature construction in a bioinformatics case study, named prediction of protein-peptide binding residue. (ii). The proposed GP algorithm has been adopted to construct new high-level and powerful features through mathematical transformations according to the hidden relations of low-level features. (iii). An impressive point relies on applying an optimization-based fitness function in GP to attain all the improved combinations of used features with various natures. (iv). The potential of GP-feature construction along with individual-based (GP+SVM) [9] and ensemble-based (SPPPred) have been confirmed. (v) Consistent and robust performance for SPPPred on both the ten-fold cross-validation and independent tests have been obtained by sequence-based information. (vi) Interdisciplinary-based evaluation was performed. (vii). Predicted results demonstrated that the SPPPred surpasses other state-of-the-art methods with F-M, ACC, and MCC of 0.310, 0.949, and 0.230, respectively. In future works, we will try to employ a deep learning-based predictor by using three-dimensional structures of proteins that are available in a public database, named Alpha-Fold2 as well as biological information like conservation, homology, and biological function.

Data availability and implementation: https://github.com/GTaherzadeh/SPPPred.git

Contact: shafiee.shima@razi.ac.ir

References

[1] S. A. Chetachukwu, R. Tahergorabi, and S. V. Hosseini, "Chapter 2 - proteins, peptides, and amino acids," in *Nutraceutical and Functional Food Components (Second Edition)*, ed. Cambridge, MA, USA: Academic, 2022, pp. 19–48, doi: 10.1016/B978-0-323-85052-0.00014-3.

[2] S. Shafiee and A. Fathi, "Prediction of protein-peptide-binding amino acid residues regions using machine learning algorithms," in *Proc. IEEE 26th Int. Conf. Comput. Comput. Soc. Iran*, 2021, pp. 1–6, doi: 101109/CSICC52343.2021.9420568.

[3] M. Zhou, Q. Li, and R. Wang, "Current experimental methods for characterizing protein-protein interactions," *ChemMedChem.*, vol. 20, no. 8, pp. 738–756, 2016, doi: 10.1002/cmdc.201500495.

[4] V. Bianchi, I. Mangone, F. Ferre, M. H. Citterich, and G. Ausiello, "webPDBinder: A server for the identification of ligand binding sites on protein structures," *Nucleic Acids Res.*, vol. 41, pp. W308–W313, 2013, doi: 10.1093/nar/gkt457.

[5] T. Litfin, Y. Yang, and Y. Zhou, "Spot-peptide: Template-based prediction of peptide-binding proteins and peptide-binding sites," *J. Chem. Inf. Model.*, vol. 59, no. 2, pp. 924–930, 2019, doi: 10.1021/acs.jcim.8b00777.

[6] H. Lee and C. Seok, "Template-based prediction of protein-peptide interactions by using GalaxyPepDock," in *Modeling Peptide-Protein Interactions*, ed. Berlin, Germany: Springer, pp. 37–47, 2017, doi: 10.1007/978-1-4939-6798-8_4.

[7] I. J. Åkhe, C. Mirabello, and B. Wallner, "Predicting protein-peptide interaction sites using distant protein complexes as structural templates," *Sci. Rep.*, vol. 9, no. 1, 2019, Art. no. 4267, doi: 10.1038/s41598-019-38498-7.

[8] G. Taherzadeh, Y. Zhou, A. W. C. C. Liew, and Y. Yang, "Structure-based prediction of protein–peptide-binding regions using Random Forest," *Bioinformatics*, vol. 34, no. 3, pp. 477–484, 2017, doi: 10.1093/bioinformatics/btx614.

[9] S. Shafiee and A. Fathi, "Combination of genetic programming and support vector machine-based prediction of protein-peptide binding sites with sequence and structure-based features," *J. Comput. Secur.*, vol. 8, no. 1, pp. 45–63, 2021, doi: 10.22108/JCS.2021.126817.1062.

[10] G. Taherzadeh, Y. Yang, T. Zhang, A. W. C. Liew, and Y. Zhou, "Sequence-based prediction of protein-peptide binding sites using support vector machine," *J. Comput. Chem.*, vol. 37, no. 13, pp. 1223–1229, 2016, doi: 10.1002/jcc.24314.

[11] S. Shafiee, A. Fathi, and F. A. Mohammadi, "Prediction of protein-peptide binding residues using classification algorithms," in *Proc. IEEE 20th Int. Conf. Bioinf. Bioeng.*, 2020, pp. 29–34, doi: 10.1109/BIBE50027.2020.00013.

[12] S. Gattani, A. Mishra, and M. T. Hoque, "Sequence and structure-based protein peptide binding residue prediction," in *Proc. 6th Annu. Conf. Comput. Biol. Bioinf.*, 2018. [Online]. Available: https://www.researchgate.net/publication/324831694_Sequence_and_Structure_based_Protein_Peptide_Binding_Residue_Prediction

[13] Z. Zhao, Z. Peng, and J. Yang, "Improving sequence-based prediction of protein-peptide binding residues by introducing intrinsic disorder and a consensus method," *J. Chem. Inf. Model.*, vol. 58, no. 7, pp. 1459–1468, 2018, doi: 10.1021/acs.jcim.8b00019.

[14] Z. Qiu and X. Wang, "Improved prediction of protein ligand-binding sites using random forests," *Protein Peptide Lett.*, vol. 18, no. 12, pp. 1212–1218, 2011, doi: 10.2174/092986611797642788.

[15] S. Iqbal and M. T. Hoque, "PBRpredict-suite: A suite of models to predict peptide-recognition domain residues from protein sequence," *Bioinformatics*, vol. 34, no. 19, pp. 3289–3299, 2018, doi: 10.1093/bioinformatics/bty352.

[16] B. Q. Li, Y. H. Zhang, M. L. Jin, T. Huang, and Y. D. Cai, "Prediction of protein-peptide interactions with nearest neighbor algorithm," *Curr. Bioinf.*, vol. 13, pp. 14–24, 2018, doi: 10.2174/1574893611666160711162006.

[17] A. Sharma, E. R. Vans, D. Shigemizu, and K. A. Boroevich, "DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture," *Sci. Rep.s*, vol. 9, 2019, Art. no. 11399, doi: 10.1038/s41598-019-47765-6.

[18] A. Sharma, A. Lysenlco, K. A. Boroevich, E. R. Vans, and T. Sunoda, "Deep feature: Feature selection in nonimage data using convolutional neural network," *Brief. Bioinf.*, vol. 22, 2021, Art. no. bbab297, doi: 10.1093/bib/bbab297.

[19] Z. Sun et al., "To improve the predictions of binding residues with DNA, RNA, carbohydrate, and peptide via multiple-task deep neural networks," bioRxiv, 2020, doi: 10.1101/2020.02.11.943571.

[20] W. Wardah et al., "Predicting protein-peptide binding sites with a deep convolutional neural network," *J. Theor. Biol.*, vol. 496, no. 1, 2020, Art. no. 110278, doi: 10.1016/j.jtbi.2020.110278.

[21] Y. Lei et al., "A deep-learning framework for multi-level peptide-protein interaction prediction," *Nature Commun.*, vol. 12, no. 1, pp. 5465–5475, 2021, doi: 10.1038/s41467-021-25772-4.

[22] X. Zhao, Y. Zhang, and X. Du, "DFpin: Deep learning-based protein-binding site prediction with feature-based non - redundancy from RNA level," *Comput. Biol. Med.*, vol. 142, 2022, Art. no. 105216, doi: 10.1016/j.compbiomed.2022. 105216.

[23] O. Abdin, S. Nim, H. Wen, and P. M. Kim, "PepNN: A deep attention model for the identification of peptide binding sites," 2021. [Online]. Available: https://www.biorxiv.org/content/10.1101/2021.01.10.426132v2

[24] O. Abdin, H. Wen, and P. M. Kim, "Sequence and structure-based deep learning models for the identification of peptide binding sites," in *Proc. Mach. Learn. Struct. Biol. Workshop*, 2020. [Online]. Available: https://www.mlsb/io/papers/MLSB2020Sequence_and_stucture_based.pdf

[25] J. Liang, Y. Xue, and J. Wang, "Genetic programming-based feature construction methods for foreground object segmentation," *Eng. Appl. Artif. Intell.*, vol. 89, no. 12, 2020, Art. no. 103334, doi: 10.1016/j.engappai.2019.103334.

[26] L. A. Q. Dominguez, C. Morell, and S. Ventura, "A propositionalization method of multi-relational data based on Grammar-Guided Genetic Programming," *Expert System Appl.*, vol. 168, 2021, Art. no. 114263, doi: 10.1016/j.eswa.2020.114263.

[27] T. Hamelryck, "An amino acid has two sides: A new 2D measure provides a different view of solvent exposure," *Proteins: Struct. Function Bioinf.*, vol. 59, no. 1, pp. 38–48, 2005, doi: 10.1002/prot.20379.

[28] R. Heffernan et al., "Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins," *Bioinformatics*, vol. 32, no. 6, pp. 843–849, 2016, doi: 10.1093/bioinformatics/btv665.

[29] Y. Yang et al., "Spider2:A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks," in *Prediction of Protein Secondary Structure*, Berlin, Germany: Springer, 2017, doi: 10.1007/978-1-4939-6406-2_6.

[30] L. Nanni and S. Brahnam, "Set of approaches based on 3D structure and Position Specific Scoring Matrix for predicting DNA-binding proteins," *Bioinformatics*, vol. 35, no. 11, pp. 1844–1851, 2018, doi: 10.1093/bioinformatics/bty912.

[31] S. F. Altschul et al., "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997, doi: 10.1093/nar/25.17.3389.

[32] S. R. Dipta et al., "SEMal: Accurate protein malonylation site predictor using structural and evolutionary information," *Comput. Biol. Med.*, vol. 125, no. 8, 2020, Art. no. 104022, doi: 10.1016/j.compbiomed.2020.104022.

[33] G. Taherzadeh, Y. Zhou, A. W. C. Liew, and Y. Yang, "Sequence-based prediction of protein-carbohydrate binding sites using support vector machines," *J. Chem. Inf. Model.*, vol. 56, no. 10, pp. 2115–2122, 2016, doi: 10.1021/acs.jcim.6b00320.

[34] M. .N. Islam, S. Iqbal, A. R. Katabi, and M. T. Hoque, "A balanced secondary structure predictor," *J. Theor. Biol.*, vol. 389, pp. 60–71, 2015, doi: 10.1016/jjtbi.2015.10.015.

[35] S. A. V. Moghaddam, H. A. Sahaf, B. Xue, C. Hollitt, and M. Zhang, "An automatic feature construction method for salient object detection: A genetic programming approach," *Expert System Appl.*, vol. 186, no. 1, 2021, Art. no. 115726, doi: 10.1016/j.eswa.2021.115726.

[36] J. Lou, M. Vanhoucke, J. S. Coelho, and W. Guo, "An efficient genetic programming approach to design priority rules for resource-constrained project scheduling problem," *Expert System Appl.*, vol. 198, no. 1, 2022, Art. no. 116753, doi: 10.1016/j.eswa.2022.116753.

[37] L. Vanneschi and M. Castelli, "Soft target functional complexity reduction: A hybrid regularization method for genetic programming," *Expert System Appl.*, vol. 177, 2021, Art. no. 114929, doi: 10.1016/j.eswa.2021.114929.

[38] E. K. Burke, S. Gustafson, and G. Kendall, "Ramped half-n-half initialisation bias in GP," in *Proc. Genet. Evol. Computation Conf.*, 2003, pp. 1800–1801, doi: 10.1007/3-540-45110-2-71.

[39] C. C. Gómez, S. S. Sanz, and D. Camacho, "A review on ensemble methods and their applications to optimization problems," in *Applied Optimization and Swarm Intelligence*, Berlin, Germany: Springer, 2021, pp. 25–45, doi: 10.1007/978-981-16-0662-5_2.

[40] C. Tong, H. Wang, C. Yang, and X. Ni, "Group ensemble learning enhances the accuracy and convenience of SSVEP-based BCIs via exploiting inter-subject information," *Biomed. Signal Process. Control*, vol. 68, no. 5, 2021, Art. no. 102797.

[41] D. S. Luz, T. J. B. Lima, R. Silva, D. Magalhães, and F. H. D. Araújo, "Automatic detection metastasis in breast histopathological images based on ensemble learning and color adjustment," *Biomed. Signal Process. Control*, vol. 73, no. 3, 2022, Art. no. 103564, doi: 10.1016/j.bspc. 2022.103564.

[42] R. Polikar, "Ensemble-based systems in decision making," *IEEE Circuits Syst. Mag.*, vol. 6, no. 3, pp. 21–45, Third Quarter 2006, doi: 10.1109/MCAS.2006.1688199.

[43] I. Yang, A. Roy, and Y. Zhang, "BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions," *Nucleic Acids Res.*, vol. 41, vol. no. D1, pp. D1096–D1103, 2013, doi: 10.1093/nar/gks966.

[44] A. Biegert, C. E. Mayer, M. Remmert, J. Söding, and A. Lupas, "The MPI bioinformatics Toolkit for protein sequence analysis," *Nucleic Acids Res.*, vol. 34, pp. W335–W339, 2006, doi: 10.1093/nar/gkl217.

[45] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomic.*, vol. 21, no. 1, pp. 1–13, 2020, doi: 10.1186/s12864-019-6413-7.

[46] D. G. Altman and J. M. Bland, "Statistics notes: Diagnostic tests. 1: Sensitivity and specificity," *BMJ Clin. Res.*, vol. 308, no. 6943, 1994, Art. no. 1552, doi: 10.1136/bmj.308.6943.1552.

[47] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Mining*, vol. 41, no. 1, 2021, Art. no. 13, doi: 10.1186/s13040-021-00244-z.

[48] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 233–240, doi: 10.1145/1143844.1143874.

[49] C. S. Hong and S. Y. Choi, "ROC curve generalization and AUC," *J. Korean Data Inf. Sci. Soc.*, vol. 31, no. 4, pp. 477–488, 2020, doi: 10.7465/jkdi.2020.31.4.477.

[50] G. Gorog, "An excel program for calculating and plotting receiver-operator characteristic (ROC) curves, histograms and descriptive statistics," *Comput. Biol. Med.*, vol. 24, no. 3, pp. 167–169, 1994, doi: 10.1016/0010-4825(94)90012-4.

[51] A. J. Larner, "Accuracy of cognitive screening instruments reconsidered: Overall, balanced or unbiased accuracy?," *Neurodegenerative Dis. Manage.*, vol. 12, pp. 67–76, 2022, doi: 10.2217/nmt-2021-0049.

[52] W. M. Lee, "Supervised learning-classification using support vector machines," in *Python Machine Learning*. New York, NY, USA: Wiley, 2019, pp. 177–203, doi: 10.1002/9781119557500.

[53] G. D. Garson, *Data Analytics for the Social Sciences: Applications in R*, London, U.K.: Evanston, IL: Routledge, 2021, doi: 10.4324/9781003109396.

[54] W. M. Lee, "Supervised learning-classification using K-Nearest neighbors (KNN)," in *Python Machine Learning*, New York, NY, USA: Wiley, 2019, pp. 205–220, doi: 10.1002/9781119557500.ch9.

[55] A. A. Johnson, M. Q. Qtt, and M. Dogucu, "Navies bayes classification," in *Bayes Rules*, Boca Raton, FL, USA: CRC Press, 2022, doi: 10.1201/9780429288340-14.

[56] W. M. Lee, "Supervised learning-linear regression," in *Python Machine Learning*, New York, NY, USA: Wiley, 2019, pp. 119–149, doi: 10.1002/9781119557500.ch6.

Shima Shafiee is currently working toward the PhD degree with the Department of Computer Engineering and Information Technology, Razi University, Kermanshah, Iran. Her research interests include machine learning, bioinformatics, evolutionary computation, optimization, and pattern recognition of proteins by using learning-based systems. These days, she is a full-time collaborator with Abdolhossein Fathi's laboratory, Razi University.

Abdolhossein Fathi is an associate professor with the Department of Computer Engineering and Information Technology, Razi University, Kermanshah, Iran. His research interests include image and video processing, pattern recognition, medical image analysis, data compression, biometric analysis, and HDLs hardware modeling (VHDL). He is the author and co-author of several peer-reviewed journal and conference papers. These days, he is the manager of the Information Technology (IT) center with Razi University.

Ghazaleh Taherzadeh is an Assistant Professor with the Department of Math Physics and Computer Science, Wilkes University, Wilkes-Barre, PA. Her research interests include machine learning (ML) and pattern recognition, bioinformatics, and other artificial intelligence. She contributed in reviewing and editing several articles from journals and conferences. Recently, she performed as a postdoctoral researcher with the University of Maryland, Washington DC, in 2021.